7-16-2018

Metabarcoding Analysis of Prey Composition of the Copepod Calanus finmarchicus in Regions of the North Atlantic Ocean

Heidi Yeh
heidi.yeh@uconn.edu

Metabarcoding Analysis of Prey Composition of the Copepod *Calanus finmarchicus* in Regions of the North Atlantic Ocean

Heidi Yeh

B.A., Barnard College, Columbia University, 2014

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

At the

University of Connecticut

2018

APPROVAL PAGE

Masters of Science Thesis

Metabarcoding Analysis of Prey Composition of the Copepod *Calanus finmarchicus* in Regions of the North Atlantic Ocean

Presented by

Heidi Yeh, B.A.

Major Advisor_____
Ann Bucklin

Associate Advisor_____
Senjie Lin

Associate Advisor_____
George McManus

University of Connecticut

2018

ACKNOWLEDGEMENTS

Many people have provided support and encouragement over the course of this research project. I would like to thank my advisor, Ann Bucklin. Ann has been encouraging me in my academic endeavors from the time of my arrival in Connecticut, two years before I enrolled at UConn as a graduate student. The things I have learned at UConn have been greatly enhanced by the travel opportunities that she provided, through which I have discovered a newfound enthusiasm for public speaking and science communication. At her insistence, I have taken risks to pursue my passions, and will forever be grateful for her encouragement.

I would also like to thank Jennifer Questel and Kendra Maas for equipping me with the bioinformatics skills necessary to analyze metabarcoding data. Jennifer Questel also played a major role in the labwork for this project, along with Hayley DeHart, who assisted with the acquisition of supplies and maintenance of smooth laboratory operations. Thanks are also due to George McManus and Senjie Lin for providing helpful suggestions and advice as associate advisors.

This project was made possible by the invaluable archives of zooplankton samples collected by the EuroBASIN 2013 cruise of the Norwegian R/V *GO Sars*, for which I would like to thank the Captain and Crew. Sequencing was performed at the facilities of the UConn Microbial Analysis Resources and Services (MARS), for which I would like to thank Kendra Maas and her team of laboratory technicians. Funding for this project was provided by the University of Connecticut Department of Marine Sciences.

Lastly, I would like to thank my family and friends for supporting me, especially my husband who safeguarded my sanity throughout this whole graduate school experience.

TABLE OF CONTENTS

ABSTRACT

The copepod *Calanus finmarchicus* occupies a pivotal position in the pelagic food web of the North Atlantic Ocean. Prey choice and predator-prey dynamics at this trophic level can influence energy transfer through the ecosystem. To examine the copepod prey composition and test the hypothesis that *C. finmarchicus* exhibit spatial variation in diet across four basins of the North Atlantic, this study used metabarcoding analysis (high throughput sequencing of orthologous gene regions) to analyze the gut contents of *C. finmarchicus.* Zooplankton net samples were collected during a 2013 cruise of the R/V *GO Sars* in the Norwegian, Icelandic, Irminger, and Labrador basins. The V4 hypervariable region of 18S ribosomal RNA (rRNA) was used to detect a broad spectrum of potential prey items based on DNA extracted from gut contents of *C. finmarchicus*. Using "universal" eukaryotic primers necessitated the development of methods to limit the signal of copepod DNA relative to the DNA of its prey. Although limited in taxonomic resolution, the results provide evidence of patterns of time / space variation of major diet components of *C. finmarchicus* across the North Atlantic. Across the dataset, the following taxonomic groups were found in order of decreasing abundance: diatoms, dinoflagellates, cnidarians, ctenophores, apicomplexa, ciliates, nematodes, and others. Strong evidence of likely symbionts and parasites in the gut of *C. finmarchicus* was found. This study provides a foundation for future efforts that may utilize primers that target narrower taxonomic groups, and thereby expand the study of diet and potential prey selectivity of *C. finmarchicus.*

I. Introduction:

I.A. *Calanus finmarchicus* in the North Atlantic Foodweb:

The North Atlantic is a heterogenous, dynamic, and productive ocean ecosystem that supports many important fisheries. The calanoid copepod *Calanus finmarchicus* assumes a pivotal position within this ecosystem. The species exerts top-down pressure through its grazing behavior on marine protists to such an extent that it is thought to regulate phytoplankton blooms (Friedland et al., 2016). The omnivorous feeding behavior of *C. finmarchicus* affects multiple trophic levels; populations of phytoplankton and grazing microzooplankton are subject to grazing/predation by this species, which in turn affects the grazing rates of microzooplankton and the phytoplankton species composition (Nejstgaard et al., 2001). *C. finmarchicus* also exerts bottom-up influence through the nutritional value that this species offers to higher trophic levels, which is critical to the first-year survival of economically-important fish species such as mackerel (Jansen, 2016). Researchers working in the North Atlantic concluded that the abundance of calanoid copepods was the strongest predictor of mackerel larvae first year survivorship (Jansen, 2016). The "growth-mortality" hypothesis set forth by Anderson (1988) states that an organism's vulnerability to predation rapidly decreases with growth, so the size of a year class recruiting to a fishery can be impacted substantially by survival outcomes in the first year of life. The highly cannibalistic tendencies of mackerel larvae have been shown to be abated by higher densities of *Calanus* nauplii, their preferred food source (Fortier and Villeneuve, 1996). In the Gulf of Saint Lawrence, researchers concluded that an abundance of larger copepods, e.g., species of *Pseudocalanus* and *C. finmarchicus*, was far more likely to support a successful year class of mackerel than an abundance of small copepods (Paradis et al., 2012). *C. finmarchicus* nutritional quality can vary; lipid content—which is significantly correlated with

energy content—was observed to vary by as much as 13% between study years in the Bay of Fundy (McKinstry et al., 2013). Such variability in nutritional value can have cascading effects that impact the metabolic budgets of endangered right whale populations (Greene and Pershing; Meyer-Gutbrod et al., 2015). The diet of *C. finmarchicus* can impact the quality and abundance of the species as a food source, so it is important to understand its position in pelagic food web dynamics of the North Atlantic.

Questions of prey choice and selectivity are essential for understanding how changes in the prey field, including changes in composition due to phytoplankton blooms, may be translated to higher trophic levels in pelagic ecosystems. Understanding the dynamics of the lower trophic levels that support fisheries is an important component of ecosystem approaches to fisheries management. Link (2011) affirms that understanding of food web dynamics at every trophic level is important for the elucidation of patterns that inform ecosystem-based fishery management (EBFM). The diet of a key species like *C. finmarchicus* may also have a significant effect on nutrient cycles, as local carbon export (via sinking of fecal pellets) is dramatically increased when this species has a diatom-heavy diet (Urban-Rich, 2001). Given the importance of *C. finmarchicus* to the North Atlantic food web, many attempts have been made to determine the diet of the species via direct observation (e.g., incubation experiments in which prey input and output are monitored; Koski and Wexels Riser, 2006) and indirect observation (e.g., identifying prey items in the gut contents or fecal pellets; Nejstgaard et al., 2003). The composition of copepod gut contents is an important source of information about their feeding behaviors. Collection and immediate preservation of the specimens ensures less degradation than the remains found in feces, although the small size of prey and rapid digestion rates limit the usefulness of morphological examination. Genetic methods provide a way to overcome many of these limitations, since DNA does not

degrade as rapidly as morphological characters, especially the subtle traits used for taxonomic identification of many prey groups. Using DNA found in the gut contents of copepods may thus be an effective way to determine the *in situ* diet of *C. finmarchicus.*

I.B. Diet Studies:

Review of Genetic Methods Used to Study Diet:

Prior work on the question of copepod feeding selectivity has largely been limited to incubation studies, with analysis employing non-metabarcoding techniques. Several studies have used PCR analysis, usually with cloning of the PCR products of copepod gut contents, including studies of *Calanus* species (Haley et al., 2011; Nejstgaard et al., 2003; Vestheim et al., 2005) and other copepods (Craig et al., 2014; Hu et al., 2014). Quantitative PCR (qPCR) studies of copepod gut contents have been carried out for *Calanus* (Nejstgaard et al., 2008; Troedsson et al., 2009) and other copepods (Durbin et al., 2008, 2012). Additional studies have focused on laboratory-incubation experiments using field-collected specimens of *Acartia* (Durbin et al., 2012; Hu et al., 2014) and *Calanus* (Nejstgaard et al., 2003, 2008; Ray et al., 2016; Troedsson et al., 2009). PCR analysis of gut contents of field-caught copepods has been completed for *Pseudocalanus spp.* from the Eastern Bering Sea (Cleary et al., 2015), *Centropages typicus* from Narragansett Bay (Durbin et al., 2008), and *Calanus helgolandicus* from the mid-Atlantic Bight (Vestheim et al., 2005). The advances in our understanding of copepod diet brought about by genetic methods can be further expanded through the use of metabarcoding techniques.

Metabarcoding approaches (high throughput sequencing of gene regions for assessment of biodiversity) have the advantage of comprehensive analysis based on short DNA sequences. The

diversity of entire communities can be assessed through metabarcoding (Vargas et al., 2015), with a homogenization of multiple specimens from a single sample for analysis (Durbin et al., 2012). With an output of sequences that can number in the hundreds of thousands, metabarcoding dramatically increases the likelihood that rare sequences will be detected that would otherwise have been missed using standard PCR, cloning and sequencing techniques. The use of a "universal" primer, which can detect most members of large groups such as eukaryotes and prokaryotes, allows researchers to determine gut contents semi-quantitatively without bias. All of these characteristics render metabarcoding analysis a powerful tool for the study of diet through the characterization of gut contents.

High throughput metabarcoding methods have only recently been used to study copepod diet; for example, the diets of *Pseudocalanus* in the Eastern Bering Strait (Cleary et al., 2015), *Calanus* in Japan (Junya Hirai et al., 2017), and *Calanus* in Norway (Ray et al., 2016) have employed these techniques. The last one of these (Ray et al., 2016) was an incubation study that focused on phytoplankton nutritional quality, rather than the prey selectivity of copepods. The only previous studies of *C. finmarchicus* gut content DNA were conducted by Haley et al. (2011) using standard PCR for analysis of specimens collected from the Gulf of Maine, and Nejstgaard et al. (2003) in the South Atlantic Bight, as well as a study using qPCR for samples collected from the Norwegian Sea (Nejstgaard et al., 2008). Among all of these studies, only one has used metabarcoding techniques to study the gut contents of field-caught copepod samples (collected in Norway; Ray et al., 2016). Conducting metabarcoding studies for this species across more locations and timepoints has the potential to provide a detailed understanding of *C. finmarchicus* feeding behavior.

The importance of barcoding (Roslin and Majaneva, 2016) and metabarcoding (Bucklin, 2016; Pompanon et al., 2012) in dietary/food web studies has been emphasized in multiple papers, which call out the value of barcoding for its ability to clarify trophic interactions and food web dynamics. Metabarcoding methods enable the detection of DNA from soft-bodied organisms, despite the accelerated rate at which these organisms may be digested. For example, spider DNA was still detectable in the feces of lizards, despite the expectation that these soft-bodied organisms would be degraded too quickly and thoroughly for detection (Kartzinel and Pringle, 2015). Metabarcoding was also successfully employed in the Gulf of Maine to identify phytoplankton consumed by herring that were then consumed by puffins in the Gulf of Maine; the phytoplankton DNA sequences were still detectable in the feces of puffins (Bowser et al., 2013). This study demonstrated that metabarcoding is a powerful means for the detection of prey items in predator guts/feces, but also that the results can be confounded by secondary predation when studying food web interactions at higher trophic levels (Bowser et al., 2013). For studies in which the effect of secondary predation is a concern, comparison of predator and prey diet is required (Bowser et al., 2013; Shehzad et al., 2012). Regardless of whether the DNA that is found in the gut is the product of *C. finmarchicus* feeding behaviors—or is instead the result of secondary predation, symbiosis, and/or parasitism—a wide net should be cast to detect all possible items. As noted above, the majority of studies on copepod feeding behaviors have been conducted as laboratory incubation experiments, which may be biased or flawed in the selection of phytoplankton cultures provided to the copepods. Consequently, unexpected food preferences—such as an affinity for microzooplankton among some copepods—were only recently discovered (Campbell et al., 2016). Metabarcoding offers several unique advantages for studies of diet, potentially achieving high taxonomic resolution despite the short and degraded

state of prey DNA sequences, and is useful for comprehensive studies of *C. finmarchicus* diet in the wild.

What is Known about *C. finmarchicus* Diet?

C. finmarchicus is omnivorous, acquiring nutrition through both phytoplankton grazing and intraguild predation: the consumption of potential competitors (Ohman and Runge, 1994). This species is known to consume and exhibit a selective preference for microzooplankton (Campbell et al., 2016), including ciliates (Leiknes et al., 2014; Nejstgaard et al., 1997), diatoms (Koski and Wexels Riser, 2006; Meyer-Harms et al., 1999), and dinoflagellates (Nejstgaard et al., 1997). The species is also known to consume cryptophytes, haptophytes (Meyer-Harms et al., 1999), and rotifers (Nejstgaard et al., 1997). The diet of *C. finmarchicus* is diverse, and possibly impacted by the ever-changing ocean environment in which they reside.

Whether this species has a static diet or one that shifts in response to changes in the environment and the available planktonic assemblage has been a topic of study. An incubation experiment using samples collected in Norwegian fjords showed that *C. finmarchicus* diet selectivity did not vary across multiple sampling locations; from this it was concluded that these copepods heavily prefer diatoms and avoid dinoflagellates, and will follow the same diet regardless of local phytoplankton abundances (Koski, 2007). The conclusion of this study of the phytoplankton preferences in copepod diets confirmed the results of a previous incubation study of similar design (Koski and Wexels Riser, 2006). The current study seeks to address the question of whether this observation of an inflexible *C. finmarchicus* diet from Norwegian fjords also applies to populations of *C. finmarchicus* residing in different basins of the North Atlantic Ocean.

I.D. Environmental Variation in the North Atlantic

The North Atlantic Ocean, which supports high populations of *C. finmarchicus*, comprises multiple basins that harbor a diverse range of environmental conditions and plankton dynamics. From these North Atlantic basins the Labrador, Irminger, Icelandic, and Norwegian basins are the subject of this study, which sourced samples from a cruise that spanned all four regions: the 2013 EuroBASIN cruise of the Norwegian research vessel R/V *GO Sars* (Fig 1). The Labrador and Icelandic basins are generally cooler, due to the stronger influence of Arctic water inputs from the East Greenland current, which brings cold and relatively fresh water into the region (Fig. 2). The Irminger and Norwegian basins are warmer, due to the stronger influence of tropical inputs from the North Atlantic current, which brings water that is warm and relatively salty into the region (Fig. 2). The general pattern of sea surface temperature variation among North Atlantic basins was supported by the temperature readings made along the cruise track of the R/V *GO Sars* (Fig. 3), which spanned approximately 10 degrees Celsius in temperature range. These marked environmental differences have been correlated with significant differences in organism size and life-history between populations of *C. finmarchicus* in the Labrador and Norwegian basins (Head et al., 2013). This study seeks to examine the diet of *C. finmarchicus*, which may be a factor that is underlying the disparities among copepod populations that reside in the four basins.

I.E. Goals and Objectives of this Study:

This research seeks to illuminate the trophic relationships of *C. finmarchicus* in the North Atlantic Ocean, with a particular focus on the marine protists upon which it feeds. To achieve

this, metabarcoding methods were employed to sequence the 18S rRNA V4 hypervariable region for pooled samples of gut contents of *C. finmarchicus*. To test the hypothesis that *C. finmarchicus* diet would vary spatially and be more similar among samples from environmentally similar basins, sequences were classified in taxonomic groups, which were then used to characterize the top prey items in each geographic region.

I.F. Background and Rationale for Methods

Specimen Preservation

If preserved quickly using 95% EtOH, DNA extracted from the gut contents is amenable to PCR amplification, which can provide accurate identification of prey groups that can withstand ethanol preservation, despite ongoing stages of digestion (Durbin et al., 2012). Upon collection from the field, samples should be preserved immediately in 95% ethanol, and then washed by 3x transfer in 95% ethanol prior to analysis to remove external debris that might contaminate results (Craig et al., 2014).

Universal Primers

Primer choice determines the range of organisms that can possibly be detected using genetic methods. Primers that target particular prey groups can be designed using the methodology described by Jarman et al. (2004) and were employed by Passmore et al. (2006). Using primers that are targeted to specific prey items or groups requires some prior knowledge of diet, but have the potential advantage of avoiding interference from the DNA of known symbionts/parasites or the predator itself. In a metagenetic study of lizard diet (through fecal analysis) predator DNA was avoided by creating a primer that would amplify only arthropod sequences (Kartzinel and Pringle,

2015). In a study of flies as disease vectors in Côte d'Ivoire, which included a metagenetic component, primers were selected that would only amplify the DNA of mammals, the real target of the study (Hoffmann et al., 2016). In this way, the copepod DNA could also be avoided by targeting a narrower group of potential prey items. Previous studies of zooplankton diets have targeted specific prey items, such as the coccolithophore *Emiliana huxleyi* (Nejstgaard et al., 2003), or larger taxonomic groups such as the phylum Bacillariophyta (Passmore et al., 2006). However, these approaches have not provided a comprehensive view of copepod diet: targeted primers are chosen based on assumptions from prior knowledge of the diet that may be limited in taxonomic range or resolution, potentially missing unexpected items in the diet (Pompanon et al., 2012). "Universal" eukaryotic primers allow researchers to capture nearly the full spectrum of *C. finmarchicus* diet, casting a broad net that makes no presuppositions about what eukaryotes may or may not be found.

Although efficient and versatile, this approach of using universal primers to study all eukaryotes in the sample introduces some complications for studies of diet, as both predator and prey will be amplified by the universal eukaryotic primers. Amplification biases in multi-template PCR reactions have long been the subject of study (e.g., Polz and Cavanaugh, 1998), and are especially relevant to diet studies in which the DNA of the predator is typically in far greater abundance than that of the prey items in its gut contents. Due to greater abundance, as well as the more-intact state of the predator DNA, 90% of sequences in a diet study of leopards were from the predator (Shehzad et al., 2012). In copepods, the DNA from prey items is competing with the DNA from the predator and from symbiotic ciliates, which can both overwhelm the signal of prey sequences. PCR primers can be designed to exclude copepod sequences, which was successfully done in a standard PCR study of symbiotic ciliate diversity

(Hu et al., 2014). However, this study analyzed the full-length of the 18S rRNA gene, whereas metabarcoding studies require a shorter target sequence, usually a hypervariable region within 18S (Fig. 4). A method to overcome the predator-to-prey DNA ratio problem that is compatible with metabarcoding is the use of a blocking primer, which has been employed in studies of diets of copepods (Craig et al., 2014) and bivalves (Klymus et al., 2017), for example.

Blocking Primers

To overcome the tendency of predator DNA to be amplified instead of prey DNA in the PCR process, a third primer can be added to the reaction—in addition to universal eukaryotic primers—that is specific to the predator species. When added in excess of the universal primers, these modified primers will bind to the predator DNA, preventing the PCR elongation step and effectively blocking the predator DNA from being amplified in the PCR process. Thus, they are aptly called "blocking primers".

Efforts were made in this study to design blocking primers that would reduce amplification of predator DNA in the final PCR reaction. First, an annealing-inhibiting primer was designed, which overlaps with the homologous region that would otherwise be the annealing site for the universal primer (Vestheim et al., 2011). However, given the lack of variability in the sequence immediately abutting the universal priming sites of the 18S rRNA V4 hypervariable region, it was not possible to design a primer with high enough specificity to bind to only the predator DNA. A longer overlap region could be used if designing dual-priming oligonucleotides (DPO; Vestheim et al., 2011). Since the middle section of the V4 hypervariable region exhibits highest sequence variation, several candidate species-specific priming locations were selected to test a second type of blocking primer: extension arrest primers (Vestheim et al., 2011). These

allow extension to the point of the priming site, but the partial amplicons produced by this process are then less likely to be amplified in subsequent cycles of the PCR reaction, due to their incompleteness. Although the primers were confirmed to be species-specific, when the 3' C3-spacer modification was added to the primer, no amplification was achieved, even of the prey DNA. In a study of sunfish diet, which utilized blocking primers, only 57 of the 100 digestive tracts that were analyzed were able to be PCR-amplified (Sousa et al., 2016). Given this example, it may be reasonable to expect a success rate of 50% or less samples with amplified prey DNA, even with the use of blocking primers; however, no samples were amplified when blocking primers were added into the PCR mixtures for this study.

Target specificity and primer binding strength can be significantly improved by using peptide nucleic acids (PNA) and locked nucleic acids (LNA) instead of 3' C3-spacer modifications (Vestheim et al., 2011). These alternative primers have been successfully used to study *Calanus* diet in the Pacific (Ho et al., 2017; Yi et al., 2017), but the costs associated with the synthesis of these primers is prohibitive. Given the lack of success with affordable blocking primer strategies for this research, dissection of the copepod gut (previously, full copepods had been homogenized) was employed as a means of reducing the amount of predator DNA entering the final sample.

Foregut Dissection

Even in cases when the gut is removed to limit the amount of predator DNA introduced into the final sample for DNA extraction, a large amount of cells are sloughed off from the predator stomach during the digestion process, as confirmed by studies of mammals (Shehzad et al., 2012). Removal of the copepod foregut was considered to be preferable for diet analysis,

since the contents would likely have been less affected by enzymatic action than the contents of the hind-gut and feces. However, the integrity of prey items in the foregut is also problematic, since these tend to diffuse throughout the dissection media when the foregut wall is breached, whereas the contents of the hind-gut and feces tend to remain as an intact bolus. The challenge of copepod gut dissection has long been a topic of study (Harding, 1939). A protocol for the removal of the entire copepod digestive tract is presented by Green and Shiel (1992). Foregut removal was practiced in the Vestheim and Jarman krill diet study credited with the development of effective blocking primer protocols (Vestheim and Jarman, 2008). The dissected stomachs, when removed from the exoskeleton, were then washed in fresh ethanol and homogenized for DNA extraction; the same steps were followed for this study.

Cryptic Species Discrimination

The North Atlantic is also home to a congener of *C. finmarchicus*—*Calanus glacialis*—and the two cannot be distinguished based on morphology alone (Choquet et al., 2017). *C. glacialis* is an Arctic copepod, whose southern range overlaps with the northern range of *C. finmarchicus*, so genetic methods must be employed to confirm the species identity. Instead of incurring high costs to sequence each individual copepod, insertions and deletions in a certain locus that accumulate in the respective genomes can be used to differentiate the two species. (Smolina et al., 2014). When the same PCR primers are used to amplify the DNA of the two species using standard PCR techniques, the resulting amplicons will have sizes that can be discriminated by agarose gel electrophoresis. The G_150 primer pair is an effective marker for the differentiation of *C. finmarchicus* and *C. glacialis*, with resulting amplicons that are 131 and 161 bp long, respectively (Smolina et al., 2014).

Hypervariable Regions for Metabarcoding

The V4, V7, and V9 hypervariable regions within the small subunit 18S ribosomal RNA (Fig. 4) are commonly used as barcodes. The V2, V4, and V9 (Fig. 4) have the highest variability and are therefore the most well-suited as gene markers for biodiversity studies (Hadziavdic et al., 2014). In a study comparing the use of V4 and V9 for ciliates, V4 was found to produce pairwise distances that were significantly closer to those of the full 18S rRNA gene than did V9 (Dunthorn et al., 2012) thus facilitating data comparison across more studies. In copepods, V2, V4, and V9 have been shown to be suitable for family- and order-level studies of genetic diversity, although there are some instances of species-level discrimination, e.g., V7 can resolve species of the copepod genus *Acartia* (Wu et al., 2015). The V2 and V4 regions have been found to have the largest proportion of variable sites, while V9 has the most sites that allow taxonomic resolution at the genus level (Wu et al., 2015). However, V9 is located near the end of the 18S rRNA sequence, and its usefulness may be limited by the prevalence of incomplete sequences in reference databases that may not capture this barcoding region. After consideration of previous results and basic knowledge of the evolution of the 18S rRNA gene, the V4 region was chosen as the target gene for this study.

Metabarcoding methods have an advantage over standard PCR methods of diet analysis, because high-throughput sequencing requires only relatively short sequences for assignment of taxonomic barcodes that still achieve relatively high taxonomic resolution (Deagle et al., 2006; Nejstgaard et al., 2008; Zaidi et al., 1999). The short length of sequences used for metabarcoding also confers a second advantage in diet studies: a better ratio of prey-to-predator signal. The longer the target marker gene sequence, the more likely the prey signal will be overwhelmed by the predator DNA, due to the degradation (which includes sequence shortening) brought about by

digestion of the prey DNA. In a study of sea lion feces, <2% of prey DNA was greater than 500 bp in length, (Deagle et al., 2006) rendering its detection within the much stronger signal of the predator's DNA to be much more difficult. If the majority of the prey sequences in a sample are shorter than the hypervariable region, they are also unlikely to encompass the full barcode and are therefore less likely to be detected using the barcode primers. Using the 18S rRNA V9 hypervariable region, instead of V4 of the same, was hypothesized to be advantageous for this diet study since V9 is approximately 70 bp shorter than V4, with respective lengths of ~200 and 270 bp (Fig. 4; Stoeck et al., 2010). However, trial samples that were sequenced using V9 primers (Stoeck et al., 2010) did not provide a greater number of non-copepod sequences or significantly improved rarefaction curves. Therefore, V4 primers (Stoeck et al., 2010) which provide higher taxonomic resolution of identified barcodes than V9 (Dunthorn et al., 2012) and have been widely used to study mixed assemblages, were used for this study.

Reference Database Choice

The Protist Ribosomal Reference database (PR2) database (Guillou et al., 2013) is described by its creators (stationed at the CNRS-UPMC Station Biologique in Roscoff, France) as "a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy". Every sequence is quality-checked before incorporation into this database and assigned a ranked-taxonomy that is normalized to eight terms—a large advantage over the non-normalized ranked taxonomies of GenBank (National Institutes of Health) and SILVA (German Network for Bioinformatics Infrastructure). The most recent version (v4.10.0) of PR2 contained 175,471 reference sequences. Although the database focuses on protists, it still includes enough sequences from metazoa, land plants and macrosporic fungi to facilitate a broad range of metabarcoding studies. This can be supplemented with the addition of more sequences for taxonomic groups of

interest to create a custom database. The taxonomic focus and database size render PR2 a computationally efficient reference database for this study of *C. finmarchicus* diet.

II. Methods:

II.A. Sample Collection and Selection:

Samples were collected on the 2013 cruise of the R/V *GO Sars* (Fig. 1), which spanned the months of May and June. Sea surface temperatures (SST) during this time period were measured by the CoastWatch Program, NOAA National Environmental Satellite, Data, and Information Service (NESDIS). SST data analysis and figure preparation was done using the M_map tool box in MatLab (Ver. 15B, MathWorks, Inc., Natick, MA) by Peter H. Wiebe (Woods Hole Oceanographic Institution). The grid selected was from -60W to +7E longitude and 55 to 70N latitude, 1 month time period, data set was: SST,NOAAPOES AVHRR, GAC, 0.1 degrees, Global, Day and Night. The file used: TBAsstaSmday_20130416000000_x-60_X10_y50_Y75_nx2147483647_ny2147483647.mat

From selected stations, zooplankton were sampled at a depth interval of 0-200 m using one of two ring nets: either WP-2 with 180 μm mesh or T80 with 375 μm mesh at 6 stations along the cruise track (Fig. 3). Samples were preserved in 95% EtOH and maintained at 4 ^{o}C for long-term storage. Each sample was examined under a dissecting microscope to identify 10 adult female *C. finmarchicus*; specimens were triple-washed by transfer in 95% EtOH prior to dissection. The specimens from each sample were divided into two technical replicates of 5 specimens.

II.B. ID Verification via InDel Analysis:

To confirm taxonomic identification of the *C. finmarchicus*, and discriminate the co-occurring congeneric species *C. glacialis*, genetic markers based on Insertion-Deletion (InDel) variation were used (Smolina et al., 2014). One antenna was excised from each copepod and individual DNA extractions were performed using Qiagen DNeasy DNA extraction kit (Hilden, Germany) according to the manufacturer's instructions. PCR reactions were done using reagents from the GoTaq Flexi Reaction Buffer kit (Promega Life Science, Madison, WI). Purified DNA (15 ng) was added to each reaction with the following reagents: 5.0 µl 5x green buffer, 2.5 µl $MgCl_2$, 0.7 µl dNTPs (10 µM), 0.25 µl Taq Polymerase, and 1 µl of each primer (10 µM). G_150 primers (Smolina et al., 2014) were used in the reaction with the following PCR protocol: 1 cycle of denaturation at 95 °C for 10 min; followed by 40 cycles of 95 °C for 20 sec, 55 °C for 20 sec, and 72 °C for 25 sec; a final extension cycle at 72 °C for 20 min; and an indefinite hold at 4 °C. The PCR product was run in 2.8% MetaPhor high-resolution agarose gel (Lonza Group Ltd., Basel, Switzerland) against a 50 bp Gel Pilot molecular weight marker (Qiagen, Hilden, Germany) to distinguish between *C. finmarchicus* (amplicon size 131) and *C. glacialis* (amplicon size 161; Smolina et al., 2014).

II.C. Foregut Dissection and DNA Extraction:

Following confirmation of species identification as *C. finmarchicus*, the foregut of each copepod was dissected using sterile dissection needles on an autoclaved microscope slide under a dissecting microscope. Pools of foreguts from 5 copepods were incubated for 24 hr at 56 °C in lysis buffer with Proteinase K. Then, 82.5 µl of CTAB (Cetrimonium bromide, 10% w/v) and a

full volume of phenol:chloroform:isoamyl alcohol (25:24:1) were added to the incubated solution. The supernatant of this centrifuged solution was then treated following the kit instructions of the E.Z.N.A. Mollusc DNA (Omega Bio-Tek Inc., Norcross, GA) kit to produce purified DNA for subsequent analyses.

II.D. PCR and Sequencing:

Purified DNA (12.5 ng) was added to a mixture with the following PCR reagents (Promega GoTaq® PCR Core Systems, Madison, WI): 5.0 µl 5x colorless buffer, 2.5 µl $MgCl_2$, 0.7 µl dNTPs (10 µM), 0.25 µl Taq Polymerase, and 1 µl of each primer (10 µM). The 18S rRNA V4 hypervariable region was amplified using the Reuk454FWD1 and ReukREV3 primers (Stoeck et al., 2010) using the following PCR protocol: 1 cycle of denaturation at 95 °C for 10 min; 25 cycles of 95 °C for 30 sec, 55 °C for 30 sec and 72 °C for 30 sec; final extension cycle at 72 °C for 5 min; and an indefinite hold at 4 °C. A second-step PCR was performed to attach a matching overhang plus Illumina p5/p7 and dual indexes (Lange et al., 2014), for which the following PCR protocol was used: 1 cycle of denaturation at 95 °C for 3.5 min; 8 cycles of 95 °C for 30 sec, 50 °C for 30 sec and 72 °C for 90 sec; final extension cycle at 72 °C for 10 min. PCR products were pooled for quantification and visualization using the QIAxcel DNA Fast Analysis (Qiagen, Hilden, Germany). PCR products were normalized based on the concentration of DNA from 360-440 bp then pooled using the QIAgility liquid handling robot. The pooled PCR products were cleaned using the Gene Read Size Selection kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The cleaned pool was sequenced on the MiSeq using v2 2x250 base pair kit (Illumina, Inc., San Diego, CA).

II.F. Bioinformatics and Statistics:

OTUs (Operational Taxonomic Units) were generated and compared against the Protist Ribosomal Reference (PR2) database (Guillou et al., 2013) using the Mothur bioinformatics pipeline version 1.40.1 (Schloss et al., 2009). All code used for analysis is found and annotated in Appendix 1. The raw PR2 database was aligned using MAFFT (v7.305, Berkeley Software Distribution, Berkeley, CA). A Mothur script was executed using the computational resources of the Xanadu computing cluster of the University of Connecticut Health Center (UCHC). Low quality sequences were removed that contained any ambiguities or homopolymer runs of a length 8 bp or greater. Chimeras were identified using Uchime (Edgar et al., 2011) and removed from analysis. Classification of sequences using the reference database was performed twice: first, an early classification step was performed to remove copepod sequences as early as possible in the pipeline; second, classification of both sequences and OTU's was performed following more rigorous clustering. Distances were calculated and OTU's clustered, allowing for 1 bp of difference, which represents <1% of the average V4 hypervariable region length of 270 bp (Stoeck et al., 2010). Following quality control and taxonomic classification steps, rarefaction curves were generated and stacked bar graphs produced to highlight the contributions of certain taxonomic groups to the overall number of sequence reads. The alpha diversity of each sample was calculated in Mothur using methods for the Shannon diversity index and the inverse of the Simpson index. The beta diversity of each sample in the context of this dataset was calculated in Mothur according to measurements including the Bray-Curtis dissimilarity, the Jaccard index, and the Yue & Clayton measure of dissimilarity. Relationships among samples in different North Atlantic basins was evaluated by plotting Bray-Curtis dissimilarities in nonmetric

multidimensional space, executed in R (v3.4.1, The R Foundation for Statistical Computing) using the VEGAN package for community ecology statistics (Oksanen et al., 2017).

III. Results & Discussion:

III.A. Taxonomic groups in the diet

Conclusions about what was found to "dominate" the dataset, or specific stations, are based on the number of sequences that were detected for each taxonomic group, and so may also be skewed by the differing numbers of non-copepod sequences that were detected at each station. For example, there was an 8-fold difference in non-copepod sequence numbers between the largest (174B, Labrador Sea) and smallest (187B, Norwegian Sea) samples (Fig. 5). The relative contribution of each taxonomic group to the total of non-copepod sequences in each sample can be viewed in Figure 5. Quantitative estimates from metagenetic data have many sources of error, including differing copy numbers of the 18S rRNA gene in the various prey items. In addition, analysis of gut contents has the further difficulty that the ratio of prey items may be skewed by differential rates of digestion (Deagle et al., 2018). Thus, the number of sequences should be interpreted cautiously as representing the relative fraction of that taxonomic group in the *C. finmarchicus* diet.

Expected dietary items

a. Diatoms

Diatoms (phylum Bacillariophyta) had the greatest abundance of sequence numbers across the dataset. Diatom sequences were found in the diet at every station, and dominated one

of the stations from the Irminger Sea (166A,B). A total of 53.8% of these sequences was identified as the genus *Chaetoceros*, which was abundant (75-382 sequences per sample) in the Icelandic Sea (163A,B) and Labrador Sea samples (174A, 174B, 175A), but was a negligible component of the diet of *C. finmarchicus* in the other two basins (<31 sequences per sample). The next most abundant diatom genus, *Asterionellopsis*, represented only 15.4% of diatom sequences and was found primarily in the Irminger Sea (166A); the same is true for *Thalassiosira*, which represented 14.6% of diatom sequences.

Diatoms have been a well-documented component of the *C. finmarchicus* diet (Koski, 2007; Koski and Wexels Riser, 2006; Meyer-Harms et al., 1999; Mullin, 1963; Nejstgaard et al., 1997). They have been documented specifically consuming *Thalassiosira* and *Chaetoceros* (Koski, 2007), which were found in high abundances in this study. The high number of diatom sequences suggests that these copepod populations had a diet heavy in diatoms; because of this, *C. finmarchicus*-driven carbon flux is likely high in the North Atlantic, as diatom-heavy pellets have been shown to have significantly faster sinking rates, thus enhancing carbon export to the deep ocean (Urban-Rich, 2001). The way that the diatom composition was divided among genera was different among the basins, with *Chaetoceros* dominating the diatom sequences found in the Labrador and Icelandic basins, which were expected to be more similar to each other than to the other two basins.

b. Dinoflagellates

Dinoflagellates were the second most abundant taxonomic group, and dominated the diet in the Icelandic Sea samples (163A,B) as well as the Irminger Sea samples (187A,B). A total of 28.8% of the dinoflagellate sequences belonged to the parasitic order Syndiniales, which was

concentrated with 82% of sequences in a single sample from the Labrador Sea (174B). Other notable identified lineages, in decreasing order of sequence abundance across the dataset, were the genera *Ceratium, Gymnodinium,* and *Prorocentrum*.

Dinoflagellates are a well-documented part of the *C. finmarchicus* diet (Koski, 2007; Koski and Wexels Riser, 2006; Meyer-Harms et al., 1999; Mullin, 1963), but certain genera can also parasitize copepods. The Labrador Sea sample that had a high number of Syndiniales sequences (specifically of the genus *Hematodinium*) likely had at least one copepod that was infected by these dinoflagellates. *Hematodinium* is a documented parasite of amphipods and copepods, and is known to be a parasitic castrator (e.g. Shields, 1994). Zooplankton infected with *Hematodinium* are believed to be the reservoir and means of transmission of the parasite to economically important crustaceans (Small and Pagenkopp, 2011). Impacts of parasitism include impaired functioning of muscles, hemolymph, and hepatopancreas, as well as alteration of chitin deposition (Stentiford and Shields, 2005). Although parasitism by the related dinoflagellate *Blastodinium* has been documented in *C. finmarchicus* (Shields, 1994), infection in *C. finmarchicus* by *Hematodinium* has not yet been explicitly mentioned in the literature.

c. Ciliates

Ciliates were the sixth most abundant prey items in terms of sequence numbers across the dataset, with 55.1% of sequences identified as apostome ciliates in one Labrador Sea sample (175A). Just 11.5% of ciliate sequences belonged to another class, Spirotrichea, which were concentrated among samples from a single basin: the Icelandic Sea (163A,B). The remaining ciliate sequences were sparsely scattered among a variety of classes within the phylum or had relatively low taxonomic resolution.

Ciliates have been well-documented as part of the *C. finmarchicus* diet (Koski, 2007; Koski and Wexels Riser, 2006; Nejstgaard et al., 1997), but it is likely that the 55.1% of the ciliate sequences identified as the subclass Apostomatia were not components of the diet. Apostome ciliates are known to be copepod symbionts and invasive parasites (Guo et al., 2012), so the Labrador Sea Sample likely included at least one copepod that was carrying these parasites. Similar studies of copepod diet using 18S rRNA PCR analysis have made efforts to suppress the signal from known parasites such as apostome ciliates by designing ciliate-excluding eukaryote-common (CEEC) primers (Hu et al., 2014). However, the number of sequences identified to this ciliate subclass was not large enough to require suppression. The remaining portion of ciliate sequences (44.9%) likely represent dietary choices made by the copepods, but reduces their abundance relative to other dietary items.

Unexpected dietary items:

Dietary items that were unexpected based on past documentation of *C. finmarchicus* diet included cnidarians, ctenophores, and nematodes, which were ranked third, fourth, and seventh, in terms of sequence abundances across the whole dataset. Cnidarians detected were dominated by the Class Hydrozoa, which was prevalent in samples from the Labrador Sea and the Icelandic Sea, with smaller signals detected in the Norwegian Sea and Irminger Sea (Stn, 166). In the Pacific Ocean, *Calanus sinicus* has been found to graze on the organic particles/detritus of metazoans such as hydrozoans and ctenophores over the continental shelf waters off China, Japan and Korea, when phytoplankton levels are low (Yi et al., 2017). In these situations, where the average adult form of cnidarians and hydrozoans dwarf the size and mouth-opening of the copepod, it may be assumed that the prey items in question are either the eggs or larvae of the much larger prey, or represent organic detritus derived from the prey that was collected by the

copepod through grazing on marine snow. Regardless of the actual form in which cnidarian and ctenophore prey was ingested by the copepod, this study provides evidence that these taxa include prey items for *C. finmarchicus* in the North Atlantic. Based on the reference database used, the cnidarians were identified as *Aglaura hemistoma, Nanomia*_sp., and *Pantachogon haeckeli,* although these species-level identifications should be treated with caution. The ctenophore sequences were identified as *Mertensia ovum* and *Beroe sp.*.

Possible parasites and symbionts:

Using the V4 region of 18S rRNA for metagenetic analysis allowed detection of prey items from a wide array of eukaryotic taxonomic groups. It may be assumed that most items identified in the foregut were prey items of the copepod, but there are clearly some exceptions. Published studies (Guo et al., 2012; Shields, 1994) have documented a variety of parasites and symbionts of *C. finmarchicus,* and other copepods, which were found in most of the samples examined in this study. Apostome ciliates as well as dinoflagellates of the genus *Hematodinium* and *Blastodinium* are suspected symbionts and parasites that have previously been discussed (pages 21-22) as part of this dataset.

Nematodes (Phylum Nematoda) were the seventh most abundant taxonomic group across the dataset, with 89.5% of sequences in the Icelandic Sea samples (163A and 163B) and just 8.0% in the Norwegian Sea sample (153A). Cyclopoid copepods have been found to become infected by feeding on nematode larvae, which then served as a means of parasite transmission to fish feeding on the infected copepods (Hubbard et al., 2016; Moravec, 2009), but no such documentation for nematode infection in *Calanus* has been documented. Thus, whether the

presence of nematode in the foregut of these samples was purely dietary, or an infection in-the-making cannot be ascertained with this analysis.

Apicomplexa is a clade of parasitic alveolates that was found in high abundances in the Norwegian Sea (153B), but a relatively small number was also found in the Icelandic Sea sample (163B). The detection of parasitic apicomplexa in marine zooplankton is not without precedent: apicomplexans have been found to infect amphipods (Prokopowicz et al., 2010), other calanoid copepods, and one species of apicomplexan, *Ganymedes apsteini*, has been found to infect *C. finmarchicus* (Sano et al., 2016). If they are indeed parasites and not a dietary item, or the result of a technical error or artefact, this finding leads to new questions about the impact of such parasites on copepods. Gregarine apicomplexan infections in Antarctic krill have been known to negatively impact the microvilli of the gut lumen (Takahashi et al., 2009). Despite the finding of Apicomplexa in just a few samples, the number of sequences corresponded to the fifth highest relative abundance among taxonomic groups in the gut contents of *C. finmarchicus* across the dataset.

Whether all of these taxonomic groups were secondary parasites that had infested the prey items of the copepods cannot be determined by this means of analysis, but the confidence level is high that these sequences were not derived from external contamination since the copepod foregut was dissected, including no external parts of the copepod for DNA extraction.

III.B. Variation among North Atlantic regions

Bray-Curtis dissimilarity (Oksanen et al., 2017) analysis of variation among North Atlantic regions was based on the OTUs shared in common between samples and plotted with

non-metric multidimensional scaling (NMDS; Fig. 6). The samples from the Irminger and Icelandic Seas clustered closely by station, whereas samples from stations in the Norwegian and Labrador Seas were less closely related. When considering the potential pairings among the four basins by environmental similarities, the Labrador-Icelandic and Sea samples clustered loosely, with only slight overlap with the Norwegian-Irminger Sea clusters (Fig. 6).

The finding that the Norwegian and Irminger Sea clustered somewhat separately from the other two basins supports an hypothesis that the similarity in copepod diet will reflect the environmental similarities of the basins. In broader view, these results suggest that *C. finmarchicus* exhibit relative flexibility in diet, contrary to published studies of this species in populations in Norwegian fjords (Koski, 2007; Koski and Wexels Riser, 2006). If there had been no flexibility in the diet then samples from different North Atlantic regions would have had no significant differences. However, these samples are not likely to be strong representatives of the relative composition; the finding that technical replicates were typically significantly different from each other and did not cluster more closely than samples from other stations suggests that neither sample had enough individuals to show the full breadth of dietary choices that each population had made. If they had, then dietary analysis of each sample from a single station would yield similar results that cluster very closely. A greater number of copepods would need to be sampled in each technical replicate to compensate for the apparent heterogeneity of the diet among copepods taken from the same sample.

III.C. Sources of error in metabarcoding analysis of diet*:*

The relative sequence numbers provide a semi-quantitative evaluation of the diet of *C. finmarchicus*, affording a rough look at what prey items are likely to have dominated the *in situ* diet of the copepod. Some limitations are due to the nature of metabarcoding analysis, which is widely acknowledged to provide only limited quantitative information (Bik et al., 2012). One issue is that 18S rRNA gene copy numbers vary widely (Prokopowich et al., 2003) both within and between taxonomic groups of marine protists (Zhu et al., 2005). In addition, differential rates at which different prey may be digested will impact both absolute and relative numbers of sequences obtained. Efforts have been made to quantitatively characterize the diet of *C. finmarchicus*; researchers found that qPCR tends to underestimate the *in situ* feeding rate of copepods, because their digestion rates are much higher than that of other mesozooplankton (Nejstgaard et al., 2008). Another significant limitation is the identification of OTUs based on available reference databases. Large numbers of the sequences from each sample either could not be classified or were designated as 'unclassified eukaryotes' (Fig. 7), ranging from 46-94% of the sequences from individual samples. These sequences may represent DNA from prey items that had already been digested at the time of sampling, resulting in degraded sequence reads that prevented taxonomic identification of the most easily-digested items in the final results. Adjustments of the final sequence numbers could be made to correct for the fast digestion in copepods, as done in a previous study of the *C. finmarchicus* diet (Troedsson et al., 2009). Lacking a finalized methodology from these various efforts to correct for differing copy numbers, digestion rates, and degraded sequences, the data from this study is presented without any such adjustments.

Rarefaction curves (calculated as numbers of OTUs versus sequencing depth) showed steep slopes for each sample, and did not show saturation (approaching an asymptote; Fig. 8). This is an indication of insufficient sequencing depth to accurately detect all taxa, including rare groups. As this does not occur for any sample, it is evident that the results of this study provide an incomplete representation of *C. finmarchicus* diet at these locations. Therefore, this dataset is most useful as providing evidence concerning the most abundant items found in the gut of *C. finmarchicus*, but should not be viewed as a comprehensive summary of all possible dietary items for the species.

Certain taxonomic categories have been excluded from this diet study as a result of the primer choice. A eukaryotic marker gene region will fail to detect any bacterial components, which require different primers and protocols (Stackebrandt and Goodfellow, 1991). The diet of *C. finmarchicus* is thought to be mostly eukaryotic, with some strains of cyanobacteria being the main exception, considered a food of last-resort for these copepods (Meyer-Harms et al., 1999). Other taxonomic categories are missed due to the general inability of metabarcoding analysis to distinguish between members of the same species; whether a *C. finmarchicus* sequence was derived from the predator or a cannibalized relative cannot be determined with current techniques. Cannibalism is thought to be a common phenomenon among *Calanus* (Bonnet et al., 2004), and specifically plays a significant role in boosting adult *C. finmarchicus* populations (Eiane et al., 2002; Ohman and Hirche, 2001). Thus, taxonomic groups such as bacteria and other copepods are missing from this analysis of *C. finmarchicus* diet.

This study provided multiple insights that should be taken into consideration for future iterations of metagenetic analyses of *C. finmarchicus* gut contents. The presence of symbiont/parasite DNA did not inhibit the detection of gut content DNA, thus avoiding the concern

of standard PCR-based studies of diet. Furthermore, the sequence reads of ciliate groups that are known to be symbiotic/parasitic can easily be discarded from the data. Importantly, additional efforts are needed to address the dominance of predator sequences, which compromises the detection of all prey sequences and the resolution of subsequent taxonomic assignments.

III.D. Next Steps for this research:

Use of metagenetic methods that target particular taxonomic groups may also be useful to improve detection of rare taxa and taxonomic resolution of OTU identification, while strategically avoiding DNA from the predator. For example, the SAR supergroup includes heterokonts, alveolates, and rhizarians; primers that target SAR should miss copepods completely, thus allowing an in-depth look at a few key prey groups (Sisson et al., 2018). The subsequent improvement in rarefaction curves would allow for more rigorous comparison across samples and bolster confidence that the final results represent the actual ratio of sequences found in the gut contents.

Prey selectivity of *C. finmarchicus* can be determined by comparing results of metagenetic analysis of copepod gut contents to available counts of protists in the water column. Samples of zooplankton and phytoplankton in the size range of potential prey items were collected at the same locations and depth as the zooplankton samples collected for this study. When this dataset is secured, the ratio of taxonomic groups found in the gut contents and the water column can be used to evaluate questions of whether taxonomic groups were consumed according to their abundances, or selected for or against by the feeding behavior of *C.*

finmarchicus. Such insights may be used to understand possible drivers of the observed spatial variability in diet of *C. finmarchicus* collected at different stations and North Atlantic regions.

IV. Conclusions:

Metabarcoding analysis of gut contents revealed that diatoms were the dominant component of the diet of *C. finmarchicus* among the four North Atlantic Ocean regions sampled, with dinoflagellates comprising the second largest fraction of the diet. High numbers of sequences for the V4 hypervariable region of 18S rRNA were detected for a variety of other prey groups, including ciliates, cnidarians, ctenophores, nematodes, and apicomplexans. This method of gut content examination ensures the detection of a broad spectrum of prey, including rare items, as well as suspected parasites/symbionts. The diet of sampled copepods was diverse, varying among replicates and regions. Some support was found for the closer relationship between the Labrador and Icelandic basins vs. the Norwegian and Irminger basins—pairings which share similar environmental conditions and are therefore more likely to support similar plankton community dynamics. This study supports the finding that metabarcoding analyses using universal primers are an optimal choice—among methods that are currently available—for future studies of diet. However, there is still much room for metabarcoding studies of diet to improve over the methods chosen in this particular study; this need was demonstrated by non-saturating rarefaction curves and compositionally disparate technical replicates. Significant steps must be taken to reduce the predator-to-prey signal ratio and improve the sampling coverage to provide a truly comprehensive view of *C. finmarchicus* diet in the North Atlantic.

References:

Anderson, J.T. (1988). Review of size dependent survival during pre-recruit stages of fishes in relation to recruitment. J. Northwest Atl. Fish. Sci. 55–66.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R., and Thomas, W.K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol. Evol. *27*, 233–243.

Bonnet, D., Titelman, J., and Harris, R. (2004). Calanus the cannibal. J. Plankton Res. *26*, 937–948.

Bowser, A.K., Diamond, A.W., and Addison, J.A. (2013). From puffins to plankton: a DNA-based analysis of a seabird food chain in the northern Gulf of Maine. PLoS ONE *8*, e83152.

Bucklin, A. (2016). Metabarcoding analysis of zooplankton diversity: applications for monitoring regional and interannual variation of the pelagic assemblage of the NW Atlantic continental shelf. (Bergen, Norway).

Campbell, R.G., Ashjian, C.J., Sherr, E.B., Sherr, B.F., Lomas, M.W., Ross, C., Alatalo, P., Gelfman, C., and Keuren, D.V. (2016). Mesozooplankton grazing during spring sea-ice conditions in the eastern Bering Sea. Deep Sea Res. Part II Top. Stud. Oceanogr. *134*, 157–172.

Choquet, M.R.C.R.R., Haltebakk, M., Dhanasiri, A.K.S., Kosobokova, K.N., Smolina, I.V., Søreide, J., Svensen, C., Melle, W.R., Kwasniewski, S., Eiane, K., et al. (2017). Genetics redraws pelagic biogeography of Calanus. 5.

Cleary, A.C., Durbin, E.G., Rynearson, T.A., and Bailey, J. (2015). Feeding by Pseudocalanus copepods in the Bering Sea: Trophic linkages and a potential mechanism of niche partitioning. Deep Sea Res. Part II Top. Stud. Oceanogr.

Craig, C., Kimmerer, W.J., and Cohen, C.S. (2014). A DNA-based method for investigating feeding by copepod nauplii. J. Plankton Res. *36*, 271–275.

Curry, R., and Mauritzen, C. (2010). (Woods Hole Oceanographic Institution / Norwegian Meteorological Institute).

Deagle, B.E., Eveson, J.P., and Jarman, S.N. (2006). Quantification of damage in DNA recovered from highly degraded samples – a case study on DNA in faeces. Front. Zool. *3*, 11.

Deagle, B.E., Thomas, A.C., McInnes, J.C., Clarke, L.J., Vesterinen, E.J., Clare, E.L., Kartzinel, T.R., and Eveson, J.P. (2018). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? Mol. Ecol. *0*.

Dunthorn, M., Klier, J., Bunge, J., and Stoeck, T. (2012). Comparing the Hyper-Variable V4 and V9 Regions of the Small Subunit rDNA for Assessment of Ciliate Environmental Diversity. J. Eukaryot. Microbiol. *59*, 185–187.

Durbin, E.G., Casas, M.C., Rynearson, T.A., and Smith, D.C. (2008). Measurement of copepod predation on nauplii using qPCR of the cytochrome oxidase I gene. Mar. Biol. *153*, 699–707.

Durbin, E.G., Casas, M.C., and Rynearson, T.A. (2012). Copepod feeding and digestion rates using prey DNA and qPCR. J. Plankton Res. *34*, 72–82.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

Eiane, K., Aksnes, D.L., Ohman, M.D., Wood, S., and Martinussen, M.B. (2002). Stage-specific mortality of Calanus spp. under different predation regimes. Limnol. Oceanogr. *47*, 636–645.

Fortier, L., and Villeneuve, A. (1996). Cannibalism and predation of fish larvae by larvae of Atlantic mackerel, Scomber scombrus: trophodynamics and potential impact on recruitment. Oceanogr. Lit. Rev. *11*, 1155–1156.

Friedland, K.D., Record, N.R., Asch, R.G., Kristiansen, T., Saba, V.S., Drinkwater, K.F., Henson, S., Leaf, R.T., Morse, R.E., Johns, D.G., et al. (2016). Seasonal phytoplankton blooms in the North Atlantic linked to the overwintering strategies of copepods. Elem Sci Anth *4*.

Green, J.D., and Shiel, R.J. (1992). A dissection method for determining the gut contents of calanoid copepods. Trans. R. Soc. S. Aust. *116*, 129–132.

Greene, C.H., and Pershing, A.J. Climate and the conservation biology of North Atlantic right whales: the right whale at the wrong time? Front. Ecol. Environ. *2*, 29–34.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., et al. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. *41*, D597–D604.

Guo, Z., Liu, S., Hu, S., Li, T., Huang, Y., Liu, G., Zhang, H., and Lin, S. (2012). Prevalent Ciliate Symbiosis on Copepods: High Genetic Diversity and Wide Distribution Detected Using Small Subunit Ribosomal RNA Gene. PLoS ONE *7*.

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E.M., and Troedsson, C. (2014). Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. PLoS ONE *9*.

Haley, S.T., Juhl, A.R., Keafer, B.A., Anderson, D.M., and Dyhrman, S.T. (2011). Detecting copepod grazing on low-concentration populations of Alexandrium fundyense using PCR identification of ingested prey. J. Plankton Res. *33*, 927–936.

Harding, J.P. (1939). A Simple Instrument for Dissecting Minute Organisms. J. R. Microsc. Soc. *59*, 19–25.

Head, E.J.H., Melle, W., Pepin, P., Bagøien, E., and Broms, C. (2013). On the ecology of Calanus finmarchicus in the Subarctic North Atlantic: A comparison of population dynamics and environmental conditions in areas of the Labrador Sea-Labrador/Newfoundland Shelf and Norwegian Sea Atlantic and Coastal Waters. Prog. Oceanogr. *114*, 46–63.

Ho, T.W., Hwang, J.-S., Cheung, M.K., Kwan, H.S., and Wong, C.K. (2017). DNA-based study of the diet of the marine calanoid copepod Calanus sinicus. J. Exp. Mar. Biol. Ecol. *494*, 1–9.

Hoffmann, C., Stockhausen, M., Merkel, K., Calvignac-Spencer, S., and Leendertz, F.H. (2016). Assessing the feasibility of fly based surveillance of wildlife infectious diseases. Sci. Rep. *6*, srep37952.

Hu, S., Guo, Z., Li, T., Carpenter, E.J., Liu, S., and Lin, S. (2014). Detecting In Situ Copepod Diet Diversity Using Molecular Technique: Development of a Copepod/Symbiotic Ciliate-Excluding Eukaryote-Inclusive PCR Protocol. PLoS ONE *9*, 1–10.

Hubbard, I.M., Hill-Spanik, K.M., Knott, D., and Buron, I. de (2016). Development of Anguillicoloides crassus in a Cyclopoid Copepod from the Acanthocyclops robustus–americanus–vernalis Complex in South Carolina, U.S.A. Comp. Parasitol. *83*, 192–196.

Jansen, T. (2016). First-year survival of North East Atlantic mackerel (Scomber scombrus) from 1998 to 2012 appears to be driven by availability of Calanus, a preferred copepod prey. Fish. Oceanogr. *25*, 457–469.

Junya Hirai, Hidaka, K., Nagai, S., and Ichikawa, T. (2017). Molecular-based diet analysis of the early post-larvae of Japanese sardine Sardinops melanostictus and Pacific round herring Etrumeus teres. Mar. Ecol. Prog. Ser. *564*, 99–113.

Kartzinel, T.R., and Pringle, R.M. (2015). Molecular detection of invertebrate prey in vertebrate diets: trophic ecology of Caribbean island lizards. Mol. Ecol. Resour. *15*, 903–914.

Klymus, K.E., Marshall, N.T., and Stepien, C.A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS ONE *12*.

Koski, M. (2007). High reproduction of Calanus finmarchicus during a diatom-dominated spring bloom. Mar. Biol. *151*, 1785–1798.

Koski, M., and Wexels Riser, C. (2006). Post-bloom feeding of Calanus finmarchicus copepodites: Selection for autotrophic versus heterotrophic prey. Mar. Biol. Res. *2*, 109–119.

Lange, V., Böhme, I., Hofmann, J., Lang, K., Sauter, J., Schöne, B., Paul, P., Albrecht, V., Andreas, J.M., Baier, D.M., et al. (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics *15*, 63.

Leiknes, Ø., Striberny, A., Tokle, N.E., Olsen, Y., Vadstein, O., and Sommer, U. (2014). Feeding selectivity of Calanus finmarchicus in the Trondheimsfjord. J. Sea Res. *85*, 292–299.

Link, J.S., Bundy, A., Overholtz, W.J., Shackell, N., Manderson, J., Duplisea, D., Hare, J., Koen-Alonso, M., and Friedland, K.D. (2011). Ecosystem-based fisheries management in the Northwest Atlantic. Fish Fish. *12*, 152–170.

McKinstry, C.A.E., Westgate, A.J., and Koopman, H.N. (2013). Annual variation in the nutritional value of Stage V Calanus finmarchicus: implications for right whales and other copepod predators. Endanger. Species Res. *20*, 195–204.

Meyer-Gutbrod, E.L., Greene, C.H., Sullivan, P.J., and Pershing, A.J. (2015). Climate-associated changes in prey availability drive reproductive dynamics of the North Atlantic right whale population. Mar. Ecol. Prog. Ser. *535*, 243–258.

Meyer-Harms, B., Irigoien, X., Head, R., and Harris, R. (1999). Selective feeding on natural phytoplankton by Calanus finmarchicus before, during, and after the 1997 spring bloom in the Norwegian Sea. Limnol. Oceanogr. *44*, 154–165.

Moravec, F. (2009). Experimental studies on the development of Contracaecum rudolphii (Nematoda: Anisakidae) in copepod and fish paratenic hosts. Folia Parasitol. Praha *56*, 185–193.

Mullin, M.M. (1963). Some Factors Affecting the Feeding of Marine Copepods of the Genus Calanus1. Limnol. Oceanogr. *8*, 239–250.

Nejstgaard, J., Gismervik, I., and Solberg, P.T. (1997). Feeding and reproduction by Calanus finmarchicus, and microzooplankton grazing during mesocosm blooms of diatoms and the coccolithophore Emiliania huxleyi. Mar. Ecol. Prog. Ser. *147*, 197–217.

Nejstgaard, J.C., Naustvoll, L.-J., and Sazhin, A. (2001). Correcting for underestimation of microzooplankton grazing in bottle incubation experiments with mesozooplankton. 59-75.

Nejstgaard, J.C., Frischer, M.E., Raule, C.L., Gruebel, R., Kohlberg, K.E., and Verity, P.G. (2003). Molecular detection of algal prey in copepod guts and fecal pellets. Limnol. Oceanogr. Methods *1*, 29–38.

Nejstgaard, J.C., Frischer, M.E., Simonelli, P., Troedsson, C., Brakel, M., Adiyaman, F., Sazhin, A.F., and Artigas, L.F. (2008). Quantitative PCR to estimate copepod feeding. Mar. Biol. *153*, 565–577.

Ohman, M.D., and Hirche, H.-J. (2001). Density-dependent mortality in an oceanic copepod population. Nature *412*, 638–641.

Ohman, M.D., and Runge, J.A. (1994). Sustained fecundity when phytoplankton resources are in short supply: Omnivory by Calanus finmarchicus in the Gulf of St. Lawrence. Limnol. Oceanogr. *39*, 21–36.

Oksanen, J., Blanchet, G.F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., et al. (2017). vegan: Community Ecology Package.

Paradis, V., Sirois, P., Castonguay, M., and Plourde, S. (2012). Spatial variability in zooplankton and feeding of larval Atlantic mackerel (Scomber scombrus) in the southern Gulf of St. Lawrence. J. Plankton Res. *34*, 1064–1077.

Passmore, A.J., Jarman, S.N., Swadling, K.M., Kawaguchi, S., McMinn, A., and Nicol, S. (2006). DNA as a Dietary Biomarker in Antarctic Krill, Euphausia superba. Mar. Biotechnol. *8*, 686–696.

Polz, M.F., and Cavanaugh, C.M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. Appl. Environ. Microbiol. *64*, 3724–3730.

Pompanon, F., Deagle, B.E., Symondson, W.O.C., Brown, D.S., Jarman, S.N., and Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. Mol. Ecol. *21*, 1931–1950.

Prokopowich, C.D., Gregory, T.R., and Crease, T.J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. Genome *46*, 48–50.

Prokopowicz, A.J., Rueckert, S., Leander, B.S., Michaud, J., and Fortier, L. (2010). Parasitic infection of the hyperiid amphipod Themisto libellula in the Canadian Beaufort Sea (Arctic Ocean), with a description of Ganymedes themistos sp. n. (Apicomplexa, Eugregarinorida). Polar Biol. *33*, 1339–1350.

Ray, J.L., Althammer, J., Skaar, K.S., Simonelli, P., Larsen, A., Stoecker, D., Sazhin, A., Ijaz, U.Z., Quince, C., Nejstgaard, J.C., et al. (2016). Metabarcoding and metabolome analyses of copepod grazing reveal feeding preference and linkage to metabolite classes in dynamic microbial plankton communities. Mol. Ecol. n/a-n/a.

Roslin, T., and Majaneva, S. (2016). The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite! Genome *59*, 603–628.

Sano, M., Miyamoto, H., and Nishida, S. (2016). Thiriotia euchirellae n. sp., a new gregarine species (Apicomplexa: Eugregarinorida) from the mesopelagic copepod Euchirella rostrata in Sagami Bay, Japan. Mar. Biodivers. *46*, 753–760.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Shehzad, W., Riaz, T., Nawaz, M.A., Miquel, C., Poillot, C., Shah, S.A., Pompanon, F., Coissac, E., and Taberlet, P. (2012). Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (Prionailurus bengalensis) in Pakistan. Mol. Ecol. *21*, 1951–1965.

Shields, J.D. (1994). The parasitic dinoflagellates of marine crustaceans. Annu. Rev. Fish Dis. *4*, 241–271.

Sisson, C., Gulla-Devaney, B., Katz, L.A., and Grattepanche, J.-D. (2018). Seed bank and seasonal patterns of the eukaryotic SAR (Stramenopila, Alveolata and Rhizaria) clade in a New England vernal pool. J. Plankton Res.

Small, H.J., and Pagenkopp, K.M. (2011). Reservoirs and alternate hosts for pathogens of commercially important crustaceans: A review - ScienceDirect. J. Invertebr. Pathol. 153–164.

Smolina, I., Kollias, S., Poortvliet, M., Nielsen, T.G., Lindeque, P., Castellani, C., Møller, E.F., Blanco - Bercial, L., and Hoarau, G.G. (2014). Genome- and transcriptome-assisted development of nuclear insertion/deletion markers for Calanus species (Copepoda: Calanoida) identification.

Sousa, L.L., Xavier, R., Costa, V., Humphries, N.E., Trueman, C., Rosa, R., Sims, D.W., and Queiroz, N. (2016). DNA barcoding identifies a cosmopolitan diet in the ocean sunfish. Sci. Rep. *6*, srep28762.

Stackebrandt, E., and Goodfellow, M. (1991). Nucleic acid techniques in bacterial systematics (Wiley).

Stentiford, G.D., and Shields, J.D. (2005). A review of the parasitic dinoflagellates Hematodinium species and Hematodinium-like infections in marine crustaceans. Dis. Aquat. Organ. *Vol. 66*.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., and Richards, T.A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol. Ecol. *19*, 21–31.

Takahashi, K.T., Kawaguchi, S., and Toda, T. (2009). Observation by electron microscopy of a gregarine parasite of Antarctic krill: its histological aspects and ecological explanations. Polar Biol. *32*, 637–644.

Troedsson, C., Simonelli, P., Nägele, V., Nejstgaard, J., and Frischer, M. (2009). Quantification of copepod gut content by differential length amplification quantitative PCR (dla-qPCR). Mar. Biol. *156*, 253–259.

Urban-Rich, J. (2001). Seston effects on faecal pellet carbon concentrations from a mixed community of copepods in Balsfjord, Norway, and the Antarctic Polar Front. ICES J. Mar. Sci. *58*, 700–710.

Vargas, C. de, Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N.L., Probert, I., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science *348*, 1261605.

Vestheim, H., and Jarman, S.N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples – a case study on prey DNA in Antarctic krill stomachs.

Vestheim, H., Edvardsen, B., and Kaartvedt, S. (2005). Assessing feeding of a carnivorous copepod using species-specific PCR. Mar. Biol. *147*, 381–385.

Vestheim, H., Deagle, B.E., and Jarman, S.N. (2011). Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR. Methods Mol. Biol. Clifton NJ *687*, 265–274.

Wu, S., Xiong, J., and Yu, Y. (2015). Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. PLoS ONE *10*.

Yi, X., Huang, Y., Zhuang, Y., Chen, H., Yang, F., Wang, W., Xu, D., Liu, G., and Zhang, H. (2017). In situ diet of the copepod Calanus sinicus in coastal waters of the South Yellow Sea and the Bohai Sea. Acta Oceanol. Sin. *36*, 68–79.

Zaidi, R.H., Jaal, Z., Hawkes, N.J., Hemingway, J., and Symondson, W.O.C. (1999). Can multiple-copy sequences of prey DNA be detected amongst the gut contents of invertebrate predators? Mol. Ecol. *8*, 2081–2087.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol. Ecol. *52*, 79–92.

Figures:

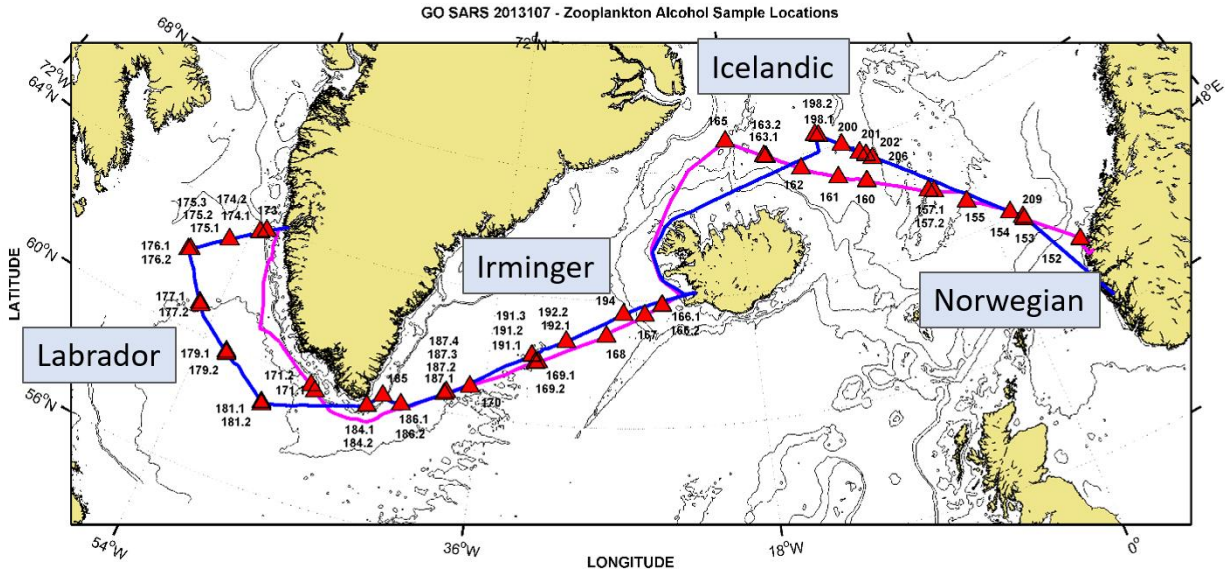

Figure 1: Cruise track of the EuroBASIN 2013 cruise of the R/V *GO Sars* with locations of stations where zooplankton samples were collected and preserved in 95% EtOH. Samples for this study were collected at stations 153, 163, 166, 174, 175, and 187. The general location of each basin mentioned throughout this paper is also labelled.

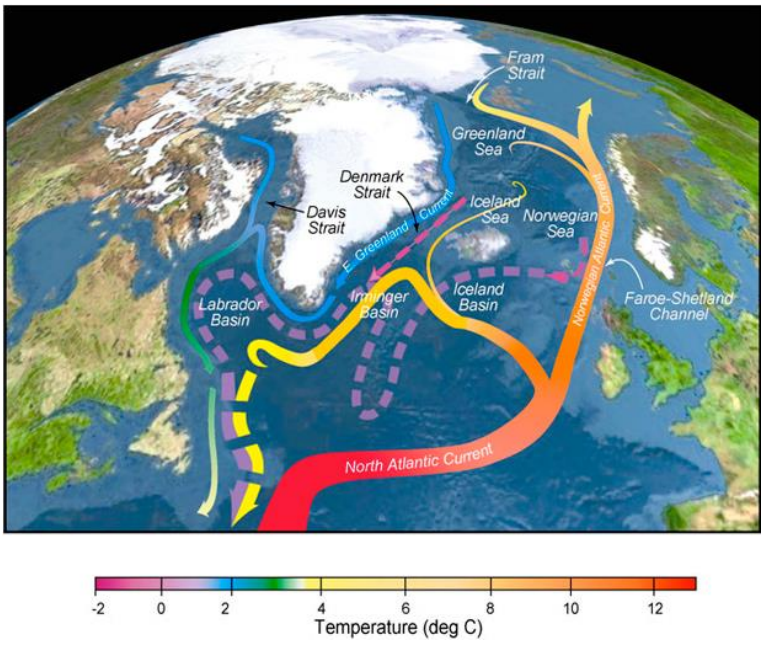

Figure 2: Map of the North Atlantic showing the major oceanic basins, as well as the temperature and relative contributions of major currents to each basin (Curry and Mauritzen, 2010).

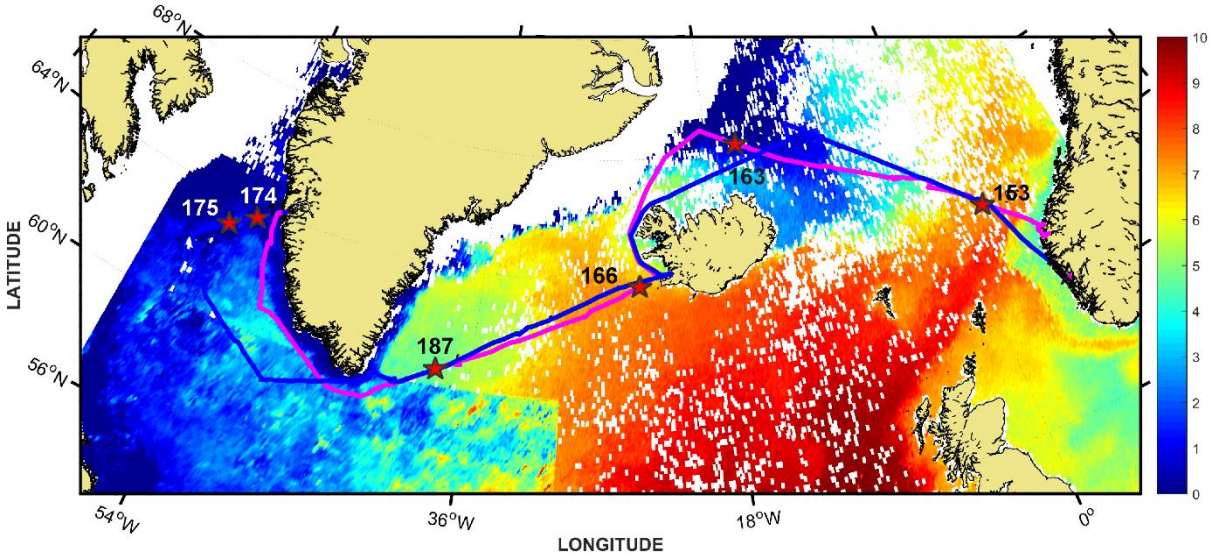

Figure 3: Sea surface temperatures across the sampled regions of the North Atlantic Ocean, including Norwegian, Icelandic, Irminger, and Labrador Seas. Lines show the cruise tracks of the EuroBASIN 2013 R/V *GO Sars* (westbound in mauve, eastbound in blue), with stars indicating stations selected for this study. Temperatures (°C) are indicated by the color bar. Data Source: CoastWatch Program, NOAA National Environmental Satellite, Data, and Information Service (NESDIS). Data analyzed and plotted by: Peter H. Wiebe, Woods Hole Oceanographic Institute.

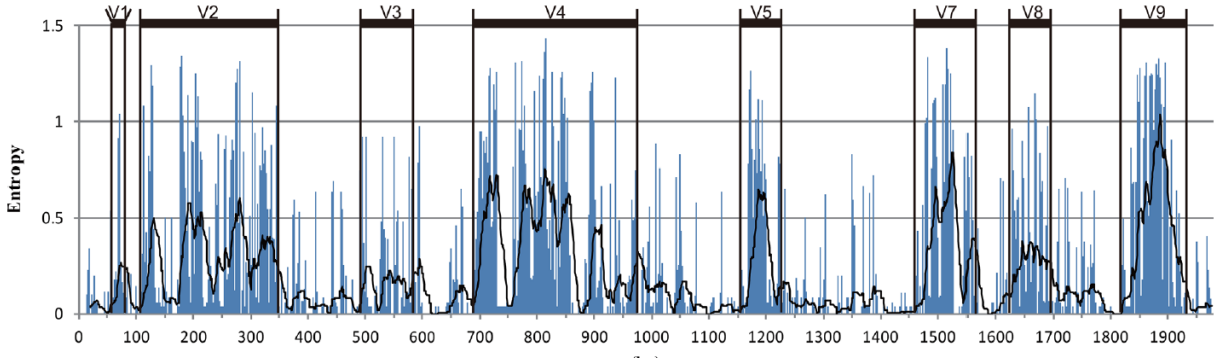

Figure 4: Copepod-specific 18S rRNA entropy (i.e., variation in DNA sequence) plot to show the distribution of V1-V9 hypervariable regions. Each point of the trendline represents the mean variability of a sliding window of 20 nucleotide base positions (Wu et al., 2015).

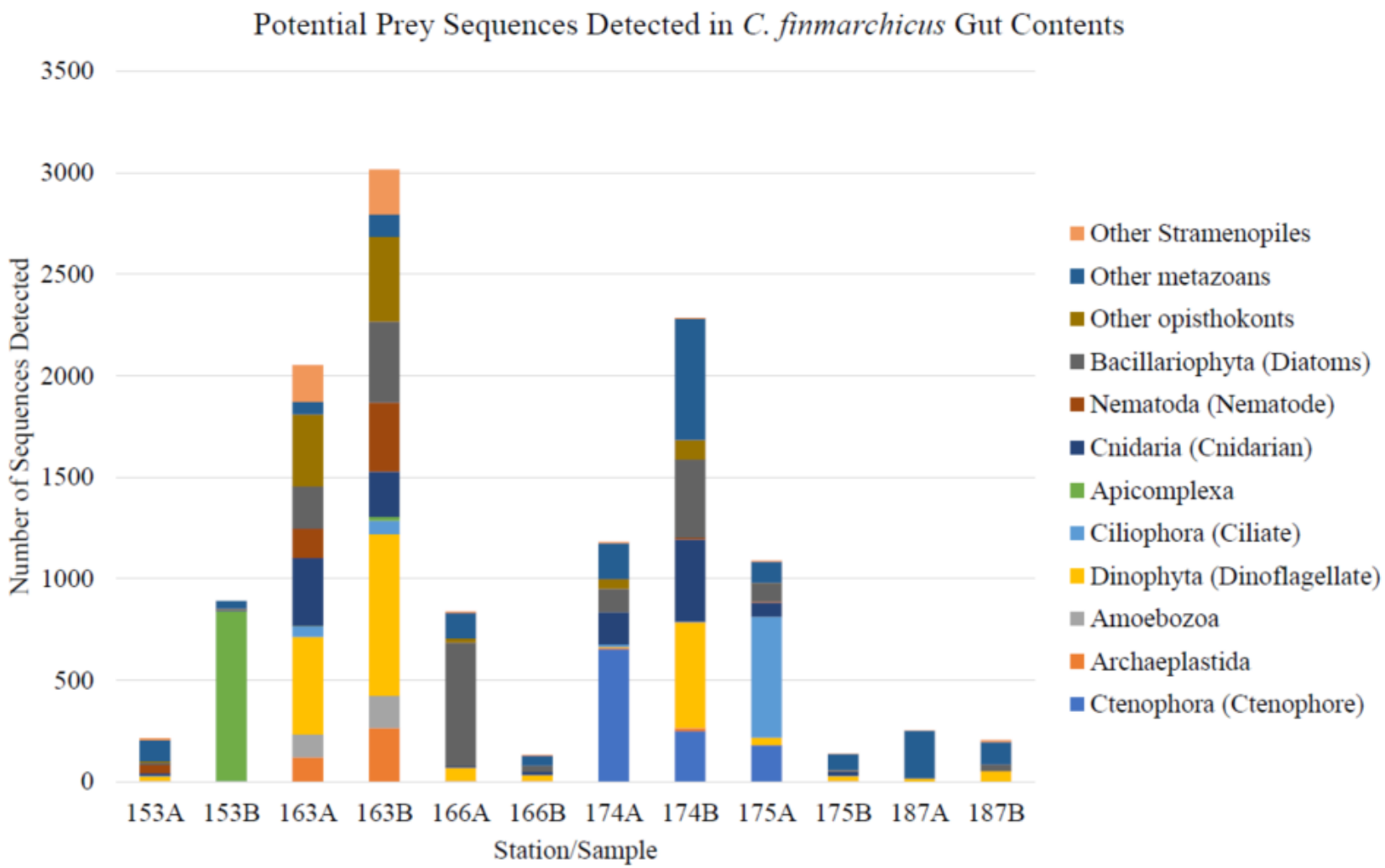

Figure 5: Absolute abundance of sequence reads classified into selected taxonomic groups. Some groups (Copepod, unknown, and unclassified/rare eukaryotes) were omitted from the analysis.

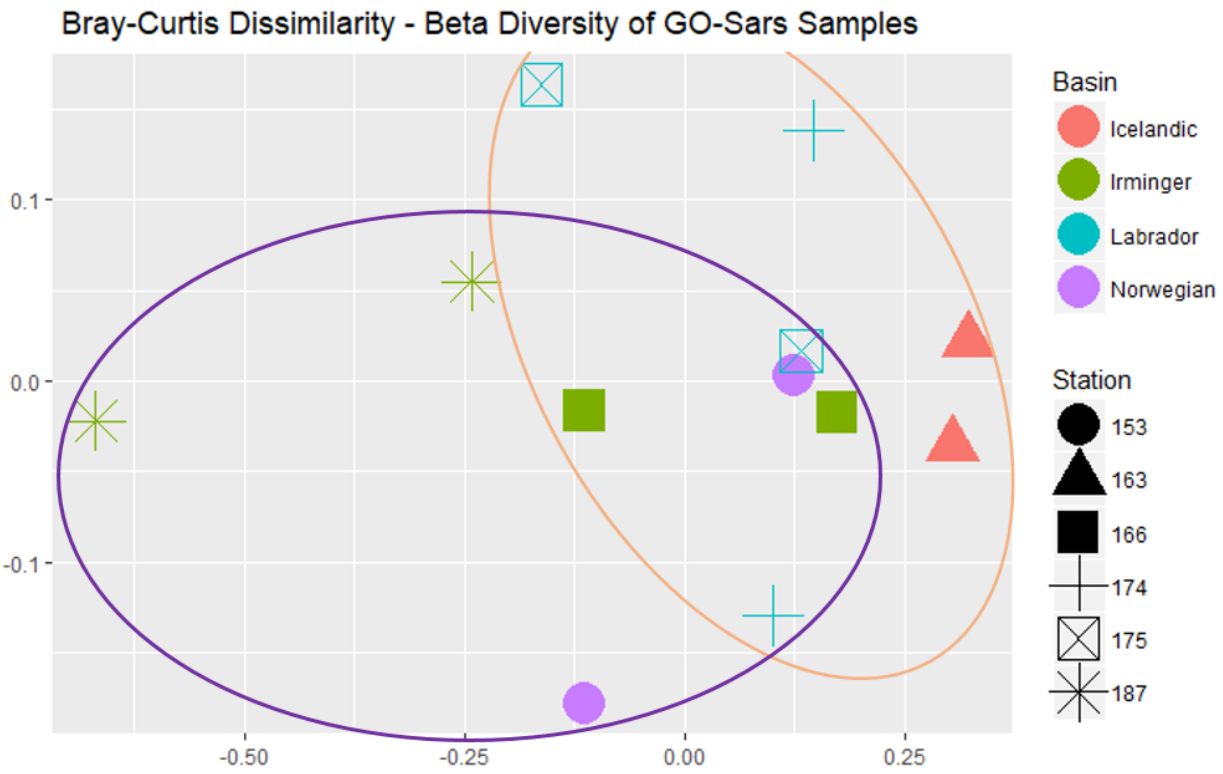

Figure 6: Bray-Curtis Dissimilarity values plotted in non-metric multidimensional scale (NMDS), in which distance represents similarity (near) or dissimilarity (far). Each station has two technical replicates, and two basins (Irminger and Labrador Seas) have biological replicates. Ovals indicate approximated clustering among environmentally-similar basins, with the purple oval representing the Irminger-Norwegian basin combination and the orange oval representing the Icelandic-Labrador basin combination.

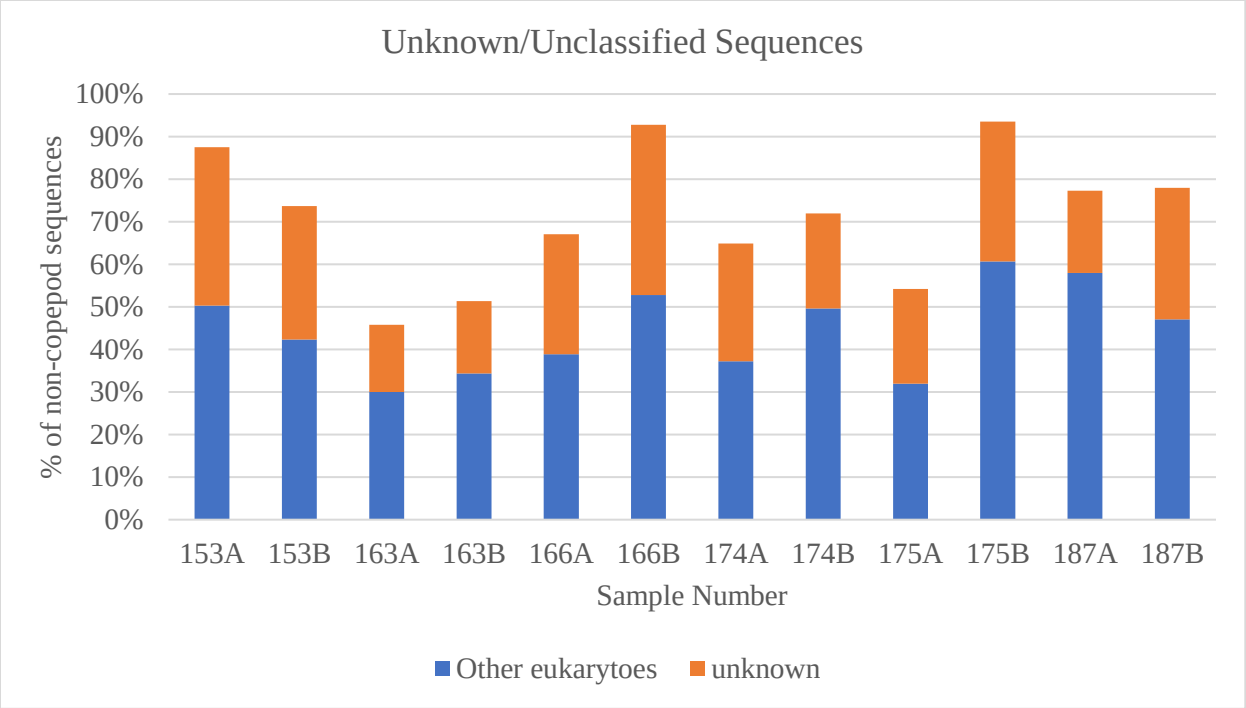

Figure 7: The percentage of non-copepod sequences from each sample that were either unknown (unclassified) or classified only 'Other eukaryotes'.

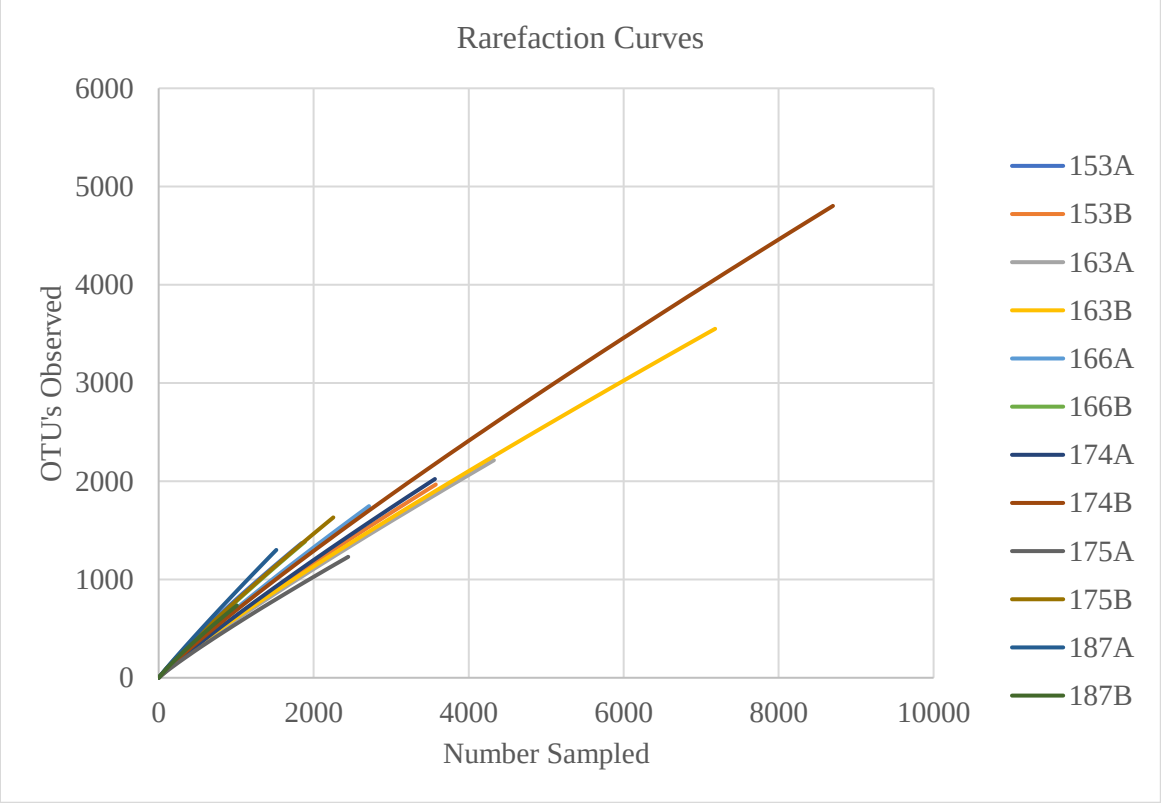

Figure 8: Rarefaction curves for each of the samples showing number of OTUs versus number of sequences or sequencing depth (Number Sampled).

Appendix 1: Annotated Mothur script

###Database Preparation

#The raw PR2 database (and any additions for a custom database) is aligned

module load mafft/7.305

mafft --retree 1 --maxiterate 0 pr2_gb203_version_4.5.fasta > Aligned_Gapped_Database.fasta

module unload

module load mothur/1.40.1

#Isolating the V4 region from 18S rRNA

pcr.seqs(fasta=pr2_gb203_version_4.5.fasta, taxonomy=pr2_gb203_version_4.5.taxo, oligos=V4.oligos, pdiffs=2, rdiffs=2)

#Creating a degapped version of this database classification steps

degap.seqs(fasta= Aligned_Gapped_Database.fasta)

###Sample Analysis

set.current(processors=48)

make.file(type=fastq, prefix=stability)

make.contigs(file=stability.files, insert=30)

summary.seqs(fasta=current)

#low-quality sequences are removed, but the usual practice of removing #sequences suspected of being too short or too long is not followed, due to #the variable sequence length of degraded gut content DNA

screen.seqs(fasta=current, group=current, summary=current, maxambig=0)

summary.seqs(fasta=current)

#A single representative is chosen for each unique sequence with multiple occurrences to reduce the computation power required for all following steps

unique.seqs(fasta=current)

summary.seqs(fasta=current, name=current)

count.seqs(name=current, group=current)

#Sample sequences are aligned against the reference database

align.seqs(fasta=current, reference=Aligned_Gapped_Database.fasta)

summary.seqs(fasta=current, count=current)

#low-quality sequences are removed, but the usual practice of removing #sequences falling outside of the 95th percentile of the alignment is not #followed, as a high abundance of predator sequences may have skewed the #alignment

screen.seqs(fasta=current, count=current, summary=current, maxhomop=8)

filter.seqs(fasta=current, vertical=T)

summary.seqs(fasta=current, count=current)

pre.cluster(fasta=current, diffs=2, count=current)

summary.seqs(fasta=current, count=current)

#Chimeras are identified and removed from subsequent analyses

chimera.uchime(fasta=current, count=current, dereplicate=t)

remove.seqs(fasta=current, accnos=current, count=current, dups=f)

summary.seqs(fasta=current, count=current)

#An early classification step is performed to remove copepod sequences from #subsequent analyses

classify.seqs(fasta=current, count=current, reference=Aligned_Degapped_Database.ng.fasta, taxonomy=pr2_gb203_version_4.5.pick.taxo)

summary.tax(taxonomy=current, count=current)

#copepod sequences are removed from analysis. A bug in this version of mothur #prevents the normal removal of a lineage string ending in a ";"

remove.lineage(fasta=current, count=current, taxonomy=current, taxon=Eukaryota;Opisthokonta;Metazoa;Arthropoda)

summary.seqs(fasta=current, count=current)

#0.006 is 1 bp difference for V4

dist.seqs(fasta=current, countends=F, cutoff=0.006)

#OTU's created via clustering to a threshold that allows one bp difference

cluster(column=current, count=current, cutoff=0.006, method=opti)

summary.seqs(fasta=current, count=current)

make.shared(list=current, count=current, label=0.006)

###check number of sequences in each sample
count.groups(shared=current)

get.current()

#Sequences and OTU's are classified via comparison to the degapped reference Edatabase

classify.seqs(fasta=current, count=current, reference=Aligned_Degapped_Database.ng.fasta, taxonomy=pr2_gb203_version_4.5.pick.taxo)

classify.otu(taxonomy=current, count=current, list=current)

#A single sequence to represent each OTU is chosen

get.outrep(fasta=current, count=current, list=current, method=abundance)

###Statistics

#alpha diversity
summary.single(shared=current, calc=nseqs-sobs-coverage-shannon-shannoneven-invsimpson)

#beta diversity
dist.shared(shared=current, calc=braycurtis-jest-thetayc)

#rarefaction curves
rarefaction.single(shared=current, calc=sobs, freq=100)